Vol. 39 No. 7 情報処理学会論文誌 July 1998

テクニカルノート

文書検索のための飾り文字画像の復元方式

天野富夫†

新聞の見出し等の情報を検索に利用するための飾り文字の復元手法について報告する．見出しには文書の内容を表す重要な情報が含まれているが，文字の背景にテクスチャを持つ，白黒が反転している等が原因で通常のOCRで処理できないことが多い．本報告では文書画像処理結果の複数候補を検索時に利用する手法を仮定して，単純な画像フィルタの組合せによって飾り文字の見出しから既存OCRで認識可能な画像を生成できることをしめす．用いたのは水平/垂直方向にOpeningを行うモルフォロジカルフィルタと白黒反転のみである．JEIDA画像データベース中の新聞紙面の見出しに本手法を適用した後，市販のOCRソフトで認識を行い有効性を検証した．

Restoration of Decorative Character Images for Document Retrieval

TOMIO AMANO†

This article describes a method for restoring decorative character images in headlines of newspapers and magazines. Although the headlines contain useful keywords for document retrieval, conventional OCRs can not recognize them because the characters are often printed in reverse and/or printed with various background texture. We made a filter which generates plural candidate images changing a small number of simple parameters (opening threshold and reversing blakc and white), so that one of the candidates contains a "normal" character image printed in black on white background. Recognizing all the candidate images to make a index, keywords in headlines are expected to be retrieved without a manual keyword entry/verify process.

1. はじめに

文書を電子化して管理・流通するシステムが普及するためには，既存の紙文書を電子化する際のコストをおさえる必要がある．たとえば，テキストをコード化して入力する際の確認・修正を省いて画像ベースで表示や要約を行うシステム[1]が提案されている．単語による検索に関しては，文書画像解析の段階ではテキストデータを一意に決定せずに，文字の切出しや認識結果の候補を含めて検索時に利用する方式[2),3)]が検討されている．

本稿では，この複数候補の考えかたによって新聞の見出し等の「飾り」のついた文字列画像を処理して検索の対象とする手法について検討する．見出しは，文書や記事の内容を短い言葉で表現しており検索のための有用な情報を含んでいる．しかし，見出し文字列は背景にテクスチャをともなったり白黒が反転して印刷されているため，通常使われているOCRでは認識できないことが多い．この問題を解決するため，OCR処理可能な画像を復元する前処理[4),5)]や飾りのついたままで認識を行うための類似度[6)]が提案されている．前者の手法には画像の復元後は既存の文字切出し・認識手法を用いてシステムを構成できるという利点がある．提案手法は，復元画像の候補を複数個生成することにより単純な前処理と既存OCR技術の組合せで検索もれの少ない頑健な文書管理システムを実現することを目指している．20個（10×2種類の解像度）の飾り文字の見出し画像から復元画像の候補を生成し，市販のOCRソフトウェアで認識させたところ良好な結果が得られた．

2. 復元画像候補の生成

図1に飾り文字見出しの復元時に複数候補を生成する場合の検索情報登録の流れをしめす．1個の飾り文字見出しから複数の復元画像を候補として出力し，それぞれに対してOCRによる文字認識処理を行った後，検索用索引に登録する．検索時に入力された文字列は

† 日本アイ・ビー・エム株式会社東京基礎研究所
Tokyo Research Laboratory, IBM Research

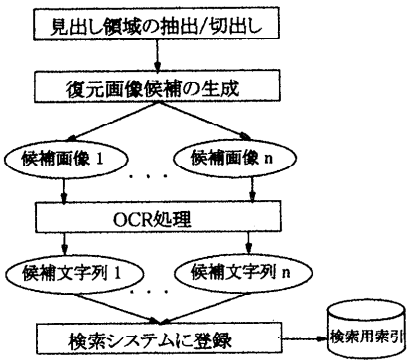
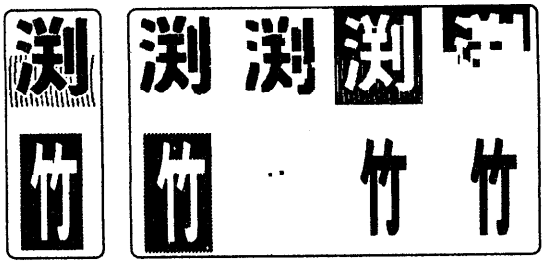


図 1 見出し情報の検索システムへの登録
Fig. 1 Registration of headlines for information retrieval.



(a) (b)

図 2 白黒反転と Opening 処理によって生成される候補画像
Fig. 2 Candidate images generated by reverse and opening. (a) Original images, (b) Generated images ($O_{thr} = 4$, $O_{thr} = 8$, Reverse & $O_{thr} = 4$, Reverse & $O_{thr} = 8$).

すべての候補の認識結果と照合される．見出し領域の抽出から検索用索引への登録までの一連の処理は，人間による確認・修正作業を経ずに自動的に行われると想定している．

飾り文字の復元処理はその前段の見出し領域の抽出結果の多少の変動——隣接する（文字の大きさや背景が異なる）2つの見出しが1つの見出し領域として抽出されるといった程度の変動——に対しては頑健である必要がある．復元処理が各見出しが1行ずつ抽出されていると仮定して，文字の幅または高さを基準として画像サイズを正規化する，線分の太さのヒストグラムから背景消去のための閾値を推定する，等々を行っていると復元は失敗し有効な検索情報が得られない．復元時に「決めうち」を避けて複数候補を生成することにより，人手による確認・修正のコストをおさえつつ

- 領域抽出結果の変動に対する頑健性が増し検索時のヒット率が向上する，
- 個々の候補の生成アルゴリズムを単純化できる（飾り文字のバリエーションに対応して生成アルゴリズムを追加していくことが容易），

といった効果が期待できる．

飾り文字画像の復元候補を生成するため次の2つの処理を行った．

(1) 文字の背景の地紋やテクスチャを消去する．
(2) 文字線分要素が黒画素のみで構成されるよう変換する．

複数候補の考えかたはどちらの処理に対しても適用可能である．(1)については1次元の structure element に基づいて Opening を行うモルフォロジカルフィルタを用意した．このフィルタは水平/垂直方向の黒画素のランを観測して長さが閾値に満たないランを消去する．水平方向の Opening は画像をラスター走査して黒画素ランを観測することに容易に行うことができる．垂直方向の Opening については画像をラスター走査しながら連続する2本の走査線のランデータを比較して縦方向に黒から白，白から黒に変わる境界線分を検出し，両者の対応をとることによって垂直方向のランの長さを求めている．一般に地紋の線の太さやテクスチャの織り目は文字本体の線の太さよりも小さいため閾値が適切に設定されていれば背景のみを消去することができる．複数の閾値で元画像を処理して候補画像を生成した．(2)について今回は白抜き文字への対処を行うこととして，上述のフィルタ処理に先立って画像の白黒を反転することによって候補画像を生成した．**図 2** に2つの処理の組合せによって生成された候補画像の例を示す．

3. 実　　験

手法の有効性を検証するため新聞紙面の見出し文字列を復元する実験を行った．実験には日本電子工業振興協会（JEIDA）作成のレイアウト解析用文書画像データベース中の新聞紙面画像（p010101-3，p010201-2）からそのままではOCRで認識できないような飾り付きの見出し10個を手作業で切り出して使用した．元画像は600 dpi の解像度でスキャンされているので1/3，1/2（200 dpi 相当と300 dpi 相当）の縮小を行い20個の画像を見出しの処理用に作成した．各画像から白黒反転処理の有無と Opening 処理の閾値（$O_{thr} = 4$ or 8）の組合せで4種類の候補画像を生成した．200 dpi 相当の画像1個からの候補生成に要した CPU 時間の平均は PC（PentiumII 266 MHz）上で約 0.23 秒であった．

結果を目視チェックしたところ，20個の見出し画像のうち18個では4種類の候補画像のいずれかに正しく復元された画像が含まれていた．200 dpi 相当の画像の復元例を**図 3** にしめす．残りの2個では一部の

図3 飾り文字画像の処理例

Fig. 3 Examples of restoration of decorative character images.

表1 OCR処理に最適な画像が得られるパラメータの分布

Table 1 Parameters giving best candidate image.

	白黒反転なし		白黒反転あり	
	$O_{thr}=4$	$O_{thr}=8$	$O_{thr}=4$	$O_{thr}=8$
200 dpi 相当	2	1	5	2
300 dpi 相当	1	2	2	5

文字列でOpeningによって文字線分が消去されてしまい完全な復元画像が得られなかった．生成する候補の数を増やしOpening処理時により小さな閾値も使うようにすれば復元の成功率はさらに高くなると期待される．

20個の見出し文字列について目視で最良の結果をしめす候補画像を市販のOCRソフトで認識（領域は手動で指定）したところ109文字について94%の認識率が得られた．**表1**に良い結果を与える候補画像がどのように分布していたかをしめす．縮小率の異なる2つのデータセットで，背景線分の太さや白抜き文字等のばらつきが4種類の候補によってカバーされていることが分かる．

4. まとめ

飾りのついた見出し文字列を検索の対象とするための画像の復元手法について検討した．復元画像の生成には見出し文字列領域の自動抽出結果に多少の誤差があることを考慮して，入力画像に対してパラメータの推定や正規化を適用しない方法を用いた．JEIDA画像データベースを用いた実験では，白黒反転の有無とOpeningの閾値2種という少数の組合せの中にOCR処理が可能なものが含まれていることが確認できた．

本手法では1個の見出し画像に対して複数の候補画像を生成するため，一意に結果を決定する場合に比較してOCRや検索エンジンが処理するデータ量は増加する．しかし，見出し1個の文字数は10から20程度なので候補生成の結果仮に100倍の量を処理することになっても，文字数の増加は通常の文書1ページ分以下である．この程度のコスト増は見出しを検索対象にできるメリットに比較して許容可能な範囲と考えられる．

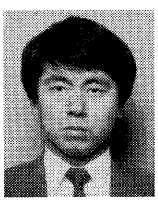
今回の実験は，背景のノイズの消去と組み合わせるフィルタとしては単純な白黒反転処理を用いた．実際の新聞紙面には白地に文字が輪郭のみで表現されている見出しもあり，このような見出しについては文献5)で行っているような背景の塗りつぶし処理も必要になる．今後はフィルタの機能拡張を行ったうえで多種の飾り文字の見出し画像に対する本手法の適用可能性を検証していきたい．

謝辞 JEIDA画像データベースを作成，配布いただいた日本電子工業振興協会認識形入力専門委員会のみなさまに感謝します．

参考文献

1) Chen, F.R. and Bloomberg, D.S.: Extraction of Indicative Summary Sentences from Imaged Documents, *ICDAR'97*, pp.227–232 (1997).
2) Senda, S., Minoh, M. and Ikeda, K.: Document Image Retrieval System Using Character Candidates Generated by Character Recognition Process, *ICDAR'93*, pp.541–546 (1993).
3) 丸川勝美，藤澤活道，嶋好 博：文書認識と全文検索の融合技術に関する実験的検討，情処情報学基礎，39-9，pp.65–71 (1995).
4) Liang, S., Ahmadi, M. and Shridhar, M.: A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background images, *CVGIP: Graphical Models and Image Processing*, Vol.56, No.5, pp.402–413 (1994).
5) 林 俊成，高井峰生，成田誠之助：画像処理による飾り文字の復元，信学技報，PRU94-12 (1994).
6) Sawaki, M. and Hagita, N.: Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning, *Trans. IEICE*, Vol.E79-D, No.5, pp.491–497 (1996).

(平成10年2月4日受付)
(平成10年5月8日採録)



天野 富夫（正会員）

昭和35年生．昭和59年慶応義塾大学大学院工学研究科計測工学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．同社東京基礎研究所にて，漢字OCR，文書画像処理の研究に従事．電子情報通信学会会員．